Predicting Critical State after COVID-19 Diagnosis Using Real-World Data from 20152 US Patients

Mike D. Rinderknecht[1*] and Yannick Klopfenstein[1*]
[1]IBM Schweiz AG, Zurich, Switzerland

The global COVID-19 pandemic caused by the virus SARS-CoV-2 has led to over 10 million confirmed cases, half a million deaths, and is challenging healthcare systems worldwide. With limited medical resources, early identification of patients with a high risk of progression to severe disease or a critical state is crucial. We present a prognostic model predicting critical state within 28 days following COVID-19 diagnosis trained on data from US electronic health records (EHR) within IBM Explorys, including demographics, comorbidities, symptoms, laboratory test results, insurance types, and hospitalization. Our entire cohort included 20152 COVID-19 cases, of which 3160 patients went into critical state or died. Random, stratified train-test splits were repeated 100 times to obtain a distribution of performance. The median and interquartile range of the areas under the receiver operating characteristic curve (ROC AUC) and the precision recall curve (PR AUC) were 0.863 [0.857, 0.866] and 0.539 [0.526, 0.550], respectively. Optimizing the decision threshold led to a sensitivity of 0.796 [0.775, 0.821] and a specificity of 0.784 [0.769, 0.805]. Good model calibration was achieved, showing only minor tendency to over-forecast probabilities above 0.6. The validity of the model was demonstrated by the interpretability analysis confirming existing evidence on major risk factors (e.g., higher age and weight, male gender, diabetes, cardiovascular disease, and chronic kidney disease). The analysis also revealed higher risk for African Americans and "self-pay patients". To the best of our knowledge, this is the largest dataset based on EHR used to create a prognosis model for COVID-19. In contrast to large-scale statistics computing odds ratios for individual risk factors, the present model combining a rich set of covariates can provide accurate personalized predictions enabling early treatment to prevent patients from progressing to a severe or critical state.

Keywords: AI, artificial intelligence, clinical decision support, coronavirus, IBM Explorys, machine learning, prognosis prediction, real-world evidence, RWE, SARS-CoV-2, triage

1 Introduction

The coronavirus disease (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Gorbalenya et al., 2020), has started to spread since the end of December 2019 from the province Hubei of the People's Republic of China to more than 180 countries becoming a global pandemic (Johns Hopkins University (JHU), 2020). Despite having a lower case fatality rate than SARS in 2003 and MERS in 2012 (Peeri et al., 2020), the overall number of 12 123 257 confirmed cases and 551 384 deaths from COVID-19 (Johns Hopkins University (JHU), 2020) (status July 9, 2020) far outweigh the other two epidemics. These high numbers have forced governments to respond with severe containment strategies to delay the spread of COVID-19 in order to avoid a global health crisis and collapse of the healthcare systems (Anderson et al., 2020; Armocida et al., 2020). Several countries have been facing shortages of intensive care beds or medical equipment such as ventilators (Ranney et al., 2020). Given these circumstances, appropriate diagnostic and prognostic tools for identifying high-risk populations and helping triage are essential for informed protection policies by policymakers and optimal allocation of resources to ensure best possible care (e.g., early treatments) for the patients.

Today's availability of data enables the development of different solutions using machine learning to address these needs, as described in the recent reviews by Bullock et al. (2020) and Wynants et al. (2020). One type of proposed solutions is prognostic prediction modeling, which consists in predicting patient outcomes such as hospitalization or exacerbation to a

Mike D. Rinderknecht and Yannick Klopfenstein are with IBM Switzerland AG, Zurich, Switzerland.

Correspondence concerning this article should be addressed to Yannick Klopfenstein, IBM Switzerland AG, Vulkanstrasse 106, 8048 Zurich, Switzerland.
E-mail: yannick.klopfenstein@ch.ibm.com

* Mike D. Rinderknecht and Yannick Klopfenstein contributed equally to this work.

critical state, using longitudinal data from medical healthcare records of COVID-19 patients (Bai et al., 2020; Feng et al., 2020; Ferrari et al., 2020; Gong et al., 2020; Haimovich et al., 2020; Jiang et al., 2020; Liu et al., 2020a; Petrilli et al., 2020; Vaid et al., 2020; Xie et al., 2020; Yan et al., 2020a) or proxy datasets based on other upper respiratory infections (DeCaprio et al., 2020). To this date, most studies include data exclusively from one or few hospitals and therefore relatively small sample sizes of confirmed COVID-19 patients (i.e., below 1000 patients), with the exception of the retrospective studies in New York City by Petrilli et al. (2020) with 4103 or by Vaid et al. (2020) with a total of 3055 patients.

The aim of this work was to create a prognostic prediction model for critical state after COVID-19 diagnosis based on a retrospective analysis of a large set of de-identified electronic health records (EHRs) of patients across the US using the IBM® Explorys® database (IBM, Armonk, NY). Such a predictive model allows identifying patients at risk based on predictive factors to support risk stratification and enable early triage.

2 Methods

2.1 RWE Insights Platform

This work was achieved by using the *RWE Insights Platform*, a data science platform for analyses of medical real-world data to generate real-world evidence (RWE) recently developed by IBM. The *RWE Insights Platform* is a data science pipeline facilitating the setup, execution, and reporting of analyses of medical real-world data to discover RWE insights in an accelerated way. The platform architecture is built in a fully modular way to be scalable to include different types of analyses (e.g., treatment pathway analysis, treatment response predictor analysis, comorbidity development analysis) and interface with different data sources (e.g., the Explorys database).

For the present use case of COVID-19 prognosis prediction, we used the comorbidity development analysis which allows defining a cohort, an outcome to be predicted, a set of predictors, and relative time windows for the extraction of the samples from the data source. New data-extraction modules for specific disease, outcome, treatments, and variables for the current use case were developed.

The *RWE Insights Platform* has been developed using open-source tools and includes a front end based on HTML and CSS interfacing via a Flask RESTful API to a Python back end (python 3.6.7) using the following main libraries: imbalanced-learn 0.6.2, numpy 1.15.4, pandas 0.23.4, scikit-learn 0.20.1, scipy 1.1.0, shap 0.35.0, statsmodel 0.90.0, and xgboost 0.90. The platform is a proprietary software owned by IBM. The detailed description of the *RWE Insights Platform* is beyond the scope of this publication.

2.2 Real-world data source

Our work was based on de-identified data from the Explorys database. The Explorys database is one of the largest clinical datasets in the world containing EHRs of around 64 million patients and spanning over 360 hospitals across the US as well as over 920 000 providers (Watson Health, IBM Corporation, 2016). Data were standardised and normalised using common ontologies, searchable through a Health Insurance Portability and Accountability Act (HIPAA)-enabled, de-identified dataset from IBM Explorys. Individuals were seen in multiple primary and secondary healthcare systems from 1999 to 2020 with a combination of data from clinical electronic medical records, health-care system outgoing bills, and adjudicated payer claims. The de-identified EHR data include patient demographics, diagnoses, procedures, prescribed drugs, vitals, and laboratory test results. Hundreds of billions of clinical, operational, and financial data elements are processed, mapped, and classified into common standards (e.g., ICD, SNOMED, LOINC, and RxNorm) within the data lake. As Explorys is updated continuously, a view of the database was created and frozen on July 16, 2020 for reproducibility of this work.

2.2.1 Cohort

The cohort included all patients in the Explorys database with a diagnosis of COVID-19 since December 1, 2019. As the new ICD-10 (International Classification of Diseases) code U07.1 for confirmed COVID-19 cases has been created and prereleased a couple of months after pandemic onset, hospitals may have used for early cases other already existing ICD codes related to coronavirus. The December 2019 cutoff was instituted to be consistent with the spread of COVID-19 in the US and to limit inclusion of patients who may have been diagnosed with other forms of coronavirus besides SARS-CoV-2. The ICD codes used to create the cohort are listed in **Table 1**. In case of multiple entries per patient after December 1, 2019, the first entry date was used as COVID-19 diagnosis date. In order to have enough data to extract the patient's outcome, the diagnosis date had to be at least 7 weeks before the freeze date of the database (July 16, 2020), as it may take up to 7 weeks from symptom onset to death (Wang et al., 2020b). LOINC codes for SARS-CoV-2 tests (e.g., 94500-6, 94309-2, 94502-2) (LOINC, 2020) with positive results available in the database were not used, as patients may have gone to a provider within the Explorys network to perform the test, but may have been treated in another hospital not covered by Explorys. This would generate a large number of additional subjects without known outcome and generate unreliable data for training the model. In contrast, patients having a diagnosis based on an ICD code may have a higher chance to be treated or have a follow-up in the same hospital.

Table 1
ICD-10 codes for the cohort. Patients with any diagnosis of the following ICD-10 codes after December 1, 2019 were considered in the cohort.

ICD-10 code	Description
U07.1	Confirmed COVID-19 case
B34.2	Coronavirus infection, unspecified
B97.29	Other coronavirus as the cause of diseases classified elsewhere

2.2.2 Prediction target

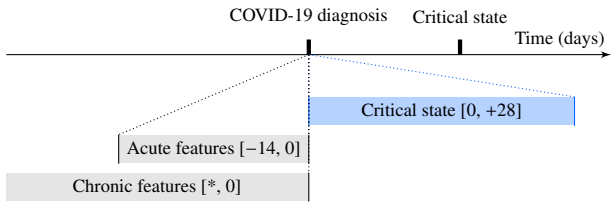
Critical state was used as a binary prediction target and included sepsis, septic shock, and respiratory failure (e.g., acute respiratory distress syndrome (ARDS)) (WHO, 2020). Severe sepsis is associated with multiple organ dysfunction syndrome. The precise definition based on ICD codes used for critical state is listed in **Table 2**. In case of multiple entries for a patient, the first entry was retained. In addition, the date of the entry for critical state had to be in a window of [0, +28] days (boundaries included) after the diagnosis date to be eligible, as illustrated in **Figure 1**). Four weeks were chosen to ensure coverage of the majority of critical outcomes, as the interquartile range of time from illness onset to sepsis and ARDS were reported to be [7, 13] and [8, 15] days, respectively (Zhou et al., 2020). Patients with an eligible entry for critical state were labeled as entering critical state, whereas patients eligible based on cohort definitions without any entry for critical state were labeled as not entering critical state. One exception to these rules were patients who are flagged as deceased in the Explorys database. In order to include death cases potentially related to COVID-19 in the critical state group, and as death dates and records with diagnoses and procedures relating to the patient's death are not available in Explorys to avoid re-identification of patients and ensure data privacy, patients with one of the following conditions were also labeled as entering critical state: deceased with an entry for critical state within the window, deceased with an entry for critical state within and after the window, or deceased without any entry for critical state (and thus excluding deceased patients with an entry for critical state before the window). In the latter case, the date was set to the end of the window for critical state entries. To validate these assumptions, the proportion of patients assumed to be deceased due to COVID-19 in our cohort was compared to epidemiological numbers.

2.2.3 Features

Features were mainly grouped into "acute" features and "chronic" features. Acute features are a set of features which should be temporally close to the COVID-19 diagnosis (e.g., recent laboratory tests, symptoms potentially related to COVID-19, or hospitalization prior to the diagnosis), whereas chronic features are a set of features which have no direct temporal relation to the COVID-19 diagnosis (e.g., chronic co-

Table 2
ICD-10 codes for the prediction target. Patients with first diagnosis of any of the listed ICD-10 codes within the specified time window were labeled as entering critical state.

ICD-10 code	Description
A41.89	Other specified sepsis
A41.9	Sepsis, unspecified organism
R65.2	Severe sepsis
R65.20	Severe sepsis without septic shock
R65.21	Severe sepsis with septic shock
J80	Acute respiratory distress syndrome (ARDS)
J96	Respiratory failure, not elsewhere classified
J96.0	Acute respiratory failure
J96.00	Acute respiratory failure, unspecified whether with hypoxia or hypercapnia
J96.01	Acute respiratory failure with hypoxia
J96.02	Acute respiratory failure with hypercapnia
J96.9	Respiratory failure, unspecified
J96.90	Respiratory failure, unspecified, unspecified whether with hypoxia or hypercapnia
J96.91	Respiratory failure, unspecified with hypoxia
J96.92	Respiratory failure, unspecified with hypercapnia



* no starting boundary

Figure 1. Time windows for prediction target and feature extraction. Schematic illustration of time window definitions relative to the COVID-19 diagnosis or to the critical state (time not to scale). The brackets define the boundaries (included) in days.

morbidities, measurable demographics, or long-term habits). Features were selected based on their appearance in literature on potential risk factors and predictors related to COVID-19. **Figure 1** illustrates their difference in terms of time windows for extraction. A negative value for boundaries of time window definitions stand for dates prior to the reference date (e.g., prior to the diagnosis date). Ideally, acute features should have been recorded for higher consistency at diagnosis date. However, this may not be always the case in the EHR compared to data from clinical studies. To account for recorded symptoms previous to the diagnosis (e.g., through tele-medicine before performing a SARS-CoV-2 test or due to potentially required multiple testing because of false negatives delaying diagnosis), a time window of [−14, 0] days before the diagnosis was used to extract acute features. Patients were considered hospitalized (inpatient) if the reported admission–discharge period of the hospitalization overlapped with the acute feature extraction time window. Entries for chronic features were considered if prior to the diagnosis date, without additional restriction. Demographic features which were not restricted to any time window (e.g., gender or race) or required a spe-

cial way of extraction/computation (e.g., age) are grouped as "special" features and are not represented in **Figure 1**. As part of the de-identification process, for patients over 90 years of age, the age is truncated to 90 years. Similarly, the age of all patients born within the last 356 days is be set to 0 years. The full list of features including their definitions (e.g., based on ICD or LOINC codes) is provided in **Table 3**, grouped by extraction time window type. As features entries (especially relevant for chronic features) may have been entered several years ago, ICD-9 codes were used as well for the extraction. In general, the last entry within the specific extraction time window was used to construct the feature, except if described otherwise in **Table 3**.

2.3 Dataset preparation and modeling approach

The full dataset was constructed based on COVID-19 diagnosis including binary prediction target labels for critical state and enriched by the various features. Patients with missing age or gender information were removed from the dataset, and all missing binary features (i.e., obtained from ICD code entries) of **Table 3** were imputed with zero. Descriptive distribution statistics were created for all features, and features with more than 90% missing values were removed from the feature set. For the remaining feature set, the concurvity (non-linear collinearity) among features was assessed using Kendall's τ, a non-parametric measure of correlation. In case of $|\tau| > 0.7$ (Dormann et al., 2013), the feature with more missing values was removed from the feature set. In case of equal number of missing values, the feature with the higher mean was removed in order to keep the minorities and make the larger group part of the predicted probability baseline. To train and evaluate the model, the dataset was split into a train set (80%) and test set (20%) using stratification of the prediction target. This procedure was repeated 100 times based on different random seeds to get a distribution and confidence intervals of the model performance and feature importance, as performance may change depending on the choice of splits.

For each random split the following steps were executed: The non-binary features of the train set and the test set were imputed based on the feature medians of the train set to avoid data leaking. An XGBoost model was trained on the train set using default parameters of the XGBoost Python package without additional hyperparameter tuning. XGBoost is a decision-tree-based ensemble machine learning algorithm using a gradient boosting framework. Gradient tree boosting models have shown to outperform other types of models on a large set of benchmarking datasets (Olson et al., 2018). The trained XGBoost model was subsequently used to create predictions for the test set.

2.4 Performance analysis and model interpretability

The performance of the model was evaluated on the test set for each random train-test split seed and reported with median and interquartile range across seeds. This provides a distribution of expected performance, if a new model would be trained on similar data. Following metrics were computed: receiver operating characteristic (ROC) curve and precision recall (PR) curve as well as their respective areas under the curve (ROC AUC and PR AUC). The confusion matrix, sensitivity, and specificity were reported for the optimal probability classification threshold. This threshold was obtained based on maximizing the largest Youden's J statistic (corresponding to the largest geometric mean as a metric for imbalanced classification seeking for a balance between sensitivity and specificity). Furthermore, the calibration of the model was reported, comparing binned mean predicted values (i.e., probabilities) to the actual fraction of positives (labeled as critical state) (Van Calster et al., 2019), in order to evaluate whether the predicted probability is realistic and can provide some confidence on the prediction.

Interpretability of the model was generated using Tree SHAP (Lundberg et al., 2020), a version of SHAP (SHapley Additive exPlanations) optimized for tree-based models. SHAP is a framework to explain the contribution of feature values to the output of individual predictions by any type of model and to compute the global importance of features. This individual contribution is expressed as SHAP value, corresponding to log-odds (output of the trees in XGBoost), before they are converted into probabilities with a logistic function. The global feature importance as well as a summary plot of individual contributions including feature values were created. In our case, a positive SHAP value indicates a contribution towards increased probability for critical state, whereas a negative SHAP value indicates a reduction of probability for critical state.

3 Results

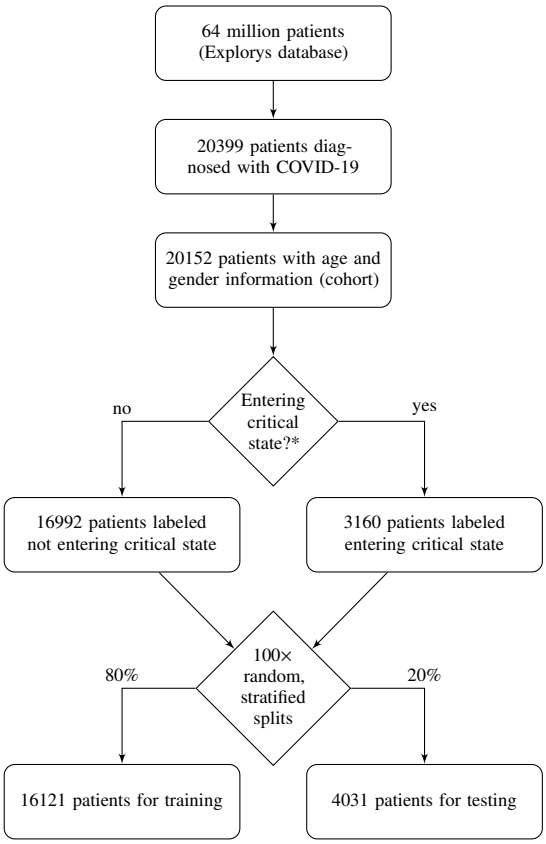
3.1 Cohort, descriptive statistics, and concurvity

The total number of identified patients diagnosed with COVID-19, the number of patients with age and gender information (referred to as the cohort), the number of patients labeled as not entering critical state and labeled as entering critical state as well as the sizes of the partitions for training and testing are reported in the schematic in **Figure 2**. Among patients labeled as critical state, a total of 1009 patients were flagged as deceased in the Explorys database. This corresponds to 5.0% of the entire cohort. **Figure 3** shows the distribution of included Explorys patients with COVID-19 diagnosis across the US. The majority of the patients are in the states LA (43.9%), OH (25.8%), DC (7.5%), FL (7.3%), and MD (7.3%). In comparison, the percentages of totally recorded patients (i.e., also non-COVID cases) in Explorys

Table 3
Feature definitions. Feature names, units and details (e.g., ICD and LOINC codes) grouped by extraction time window specifications.

Extraction time window	Feature	Units	Details
Special features	Age	Years	Computed at diagnosis date, based on birth year entry
	Gender	NA (0: male, 1: female)	No time window restrictions
	Ethnicity (Hispanic)	NA (binary)	No time window restrictions
	Ethnicity (non-Hispanic)	NA (binary)	No time window restrictions
	Ethnicity (Other)	NA (binary)	No time window restrictions
	Insurance (Medicaid)	NA (binary)	No time window restrictions
	Insurance (Medicare)	NA (binary)	No time window restrictions
	Insurance (Other)	NA (binary)	No time window restrictions
	Insurance (Other public)	NA (binary)	No time window restrictions
	Insurance (Private)	NA (binary)	No time window restrictions
	Insurance (Selfpay)	NA (binary)	No time window restrictions
	Race (African American)	NA (binary)	No time window restrictions
	Race (Asian)	NA (binary)	No time window restrictions
	Race (Caucasian)	NA (binary)	No time window restrictions
	Race (Multi-racial)	NA (binary)	No time window restrictions
	Race (Other)	NA (binary)	No time window restrictions
Acute features	Acute bronchitis	NA (binary)	ICD-10: J20.*, J40 and ICD-9: 466.0, 490
	Anorexia	NA (binary)	ICD-10: R63.0, R63.8 and ICD-9: 783.0, 783.9
	Body temperature	°C	LOINC: 8310-5
	C-reactive protein	mg/L	LOINC: 1988-5
	C-reactive protein (high sensitivity method)	mg/L	LOINC: 30522-7
	Confusion	NA (binary)	ICD-10: R41.0, R41.82 and ICD-9: 780.97
	Cough	NA (binary)	ICD-10: R05 and ICD-9: 786.2
	Diarrhea	NA (binary)	ICD-10: R19.7 and ICD-9: 787.91
	Fatigue	NA (binary)	ICD-10: R53.1, R53.81, R53.83 and ICD-9: 780.79
	Fever	NA (binary)	ICD-10: R50.9 and ICD-9: 780.60
	Headache	NA (binary)	ICD-10: R51 and ICD-9: 784.0
	Hemoptysis	NA (binary)	ICD-10: R04.2 and ICD-9: 786.30
	Hospitalization (inpatient)	NA (binary)	Considered if reported admission–discharge period overlapping with extraction time window
	Lactate dehydrogenase (L>P)	U/L	LOINC: 14804-9
	Lactate dehydrogenase (P>L)	U/L	LOINC: 14805-6
	Lymphocytes (#/100 leukocytes in blood)	%	LOINC: 26478-8
	Lymphocytes (#/blood volume)	$10^3/\mu$L	LOINC: 26474-7
	Myalgia	NA (binary)	ICD-10: M79.1, M79.10, M79.11, M79.12, M79.18 and ICD-9 729.1
	Neutrophils (#/100 leukocytes in blood)	%	LOINC: 26511-6
	Neutrophils (#/blood volume)	$10^3/\mu$L	LOINC: 26499-4
	Oxygen saturation	%	LOINC: 59408-5
	Pneumonia	NA (binary)	ICD-10: J12.*, J13, J14, J15.*, J16.*, J17, J18.* and ICD-9: 480.*, 481, 482.*, 483.*, 484.*, 485, 486, 487.0, 488.01, 488.11, 488.81
	Rhinorrhea	NA (binary)	ICD-10: J34.89 and ICD-9: 478.19
	Shortness of breath	NA (binary)	ICD-10: R06.02 and ICD-9: 786.05
	Sore throat	NA (binary)	ICD-10: J02.9 and ICD-9: 462
	Sputum	NA (binary)	ICD-10: R09.3 and ICD-9: 786.4
	Systolic blood pressure	mmHg	LOINC: 8480-6
	Vomiting	NA (binary)	ICD-10: R11.10 and ICD-9: 536.2, 787.03
Chronic features	Active smoking	NA (binary)	Based on reported habit
	Asthma	NA (binary)	ICD-10: J45.* and ICD-9: 493.*
	BMI	kg/m^2	LOINC: 39156-5, or computed from weight (29463-7) and height (8302-2)
	Cardiovascular disease	NA (binary)	ICD-10: I20.*, I21.*, I25.*, I48.*, I50.*, I63.*, I65.*, I67.*, I73.* and ICD-9: 410.*, 412.*, 413.*, 414.*, 427.*, 428.*, 429.*, 433.*, 434.*, 437.*, 443.*
	Chronic kidney disease	NA (binary)	ICD-10: E10.21, E10.22, E10.29, E11.21, E11.22, E11.29, I12.0, I12.9, I13.0, I13.10, I13.11, I13.2, N04.*, N05.*, N08, N18.*, N19, N25.9 and ICD-9: 250.40, 250.41, 250.42, 250.43, 403.*, 404.*, 581.81, 581.9, 583.89, 585.*, 588.9
	Chronic obstructive pulmonary disease	NA (binary)	ICD-10: J44.* and ICD-9: 491.*, 493.2*
	Diabetes	NA (binary)	ICD-10: E10.*, E11.*, E13.* and ICD-9: 250.*
	Hypertension	NA (binary)	ICD-10: I10, I15.* and ICD-9: 401.*, 405.*
	Immunodeficiency	NA (binary)	ICD-10: B20, D80.*, D81.*, D82.*, D83.*, D84.*, D86.*, D89.* and ICD-9: 042, 279.*
	Nicotine dependence	NA (binary)	ICD-10: F17.* and ICD-9: 305.1
	Obesity	NA (binary)	ICD-10: E66.0*, E66.1, E66.2, E66.8, E66.9 and ICD-9: 278.00, 278.01, 278.03
	Paralytic syndromes	NA (binary)	ICD-10: G80.*, G81.*, G82.*, G83.* and ICD-9: 342.*, 343.*, 344.*
	Weight	kg	LOINC: 29463-7

* symbolizes a wildcard for ICD subcategory codes.



* or deceased due to COVID-19

Figure 2. Diagram of number of subjects. Cohort selection and number of patients not entering versus entering critical state based on the definitions outlined in the according sections. To train and evaluate the model, the dataset was split using stratification of the prediction target. This procedure was repeated 100 times based on random seeds to get a distribution of model performance.

for these states are: LA (4.5%), OH (24.3%), DC (1.0%), FL (5.2%), and MD (4.3%).

Descriptive statistics after zero-imputation of the binary features and before feature reduction are reported in **Table 4**. Based on these results, the following features were removed due to a too high proportion of missing data: C-reactive protein, C-reactive protein (high sensitivity method), Lactate dehydrogenase (L>P), Lactate dehydrogenase (P>L), Lymphocytes (/100 leukocytes in blood), Lymphocytes (/blood volume), Neutrophils (/100 leukocytes in blood), Neutrophils (/blood volume), and Oxygen saturation. Rank correlations across features after removing features based on the threshold for missing data is shown in the heatmap in **Figure 4**. The following feature combinations showed a strong rank correlation: {Race (African American), Race (Caucasian)} and {BMI, Weight}, from which the following features were removed due to higher proportion of missing data or higher

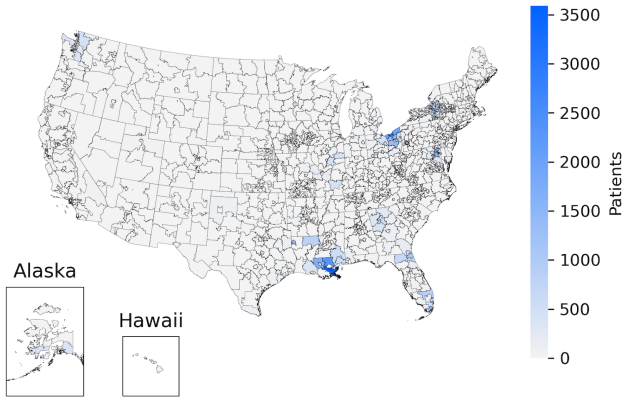


Figure 3. Distribution of cohort patients in the US. The color shade indicates the total number of COVID-19 patients for each 3-digit ZIP code of the US being recorded within the Explorys network.

mean: Race (Caucasian) and BMI.

3.2 Performance

The performance and calibration of the model was evaluated on the 4031 patients of the test set for each train-test split seed. The ROC AUC and PR AUC across different seeds were 0.863 [0.857, 0.866] and 0.539 [0.526, 0.550], respectively. **Figure 5** shows their distributions, together with the ROC curve and the precision recall curve. The confusion matrix for the identified optimal classification threshold is shown in **Figure 6**. The sensitivity of the model for this optimal threshold was 0.796 [0.775, 0.821] and the specificity 0.784 [0.769, 0.805]. The calibration of the model is shown in **Figure 7**.

3.3 Model interpretability

Figure 8 shows the results of the model interpretability analysis based on Tree SHAP. Pneumonia and older age are by far the principal predictors for critical state. The main features contributing to a higher probability of critical state in case of high feature values or presence are (in decreasing order of global feature importance): pneumonia, older age, hospitalization (inpatient), weight, shortness of breath, diabetes, race (African American), and cardiovascular disease. The main features leading to lower probability of critical state in are female gender and cough. Note that for binary features "max" feature values correspond to 1 (e.g., presence of the feature). In the case of gender, 1 corresponds to female (see **Table 3**).

4 Discussion

In this work, a prognostic model was created based on real-world data from 16121 patients to predict at COVID-19

Table 4

Descriptive statistics of the features. The descriptive statistics are based on the full dataset after zero-imputation of the binary features but before feature reduction. The percentages 25%, 50%, and 75% refer to the first (Q1), second (median), and third quartiles (Q3). Note that for binary features the Mean column represents the proportion of positive entries. Note that as part of Explorys' de-identification process the feature Age has a ceiling effect at 90 years, and the age of all patients born in the last 365 days is reported as zero. For gender, 1 corresponds to female.

Feature	Count	Missing	Mean	Std	Min	25%	50%	75%	Max
Active smoking	20152	0.0%	0.139	0.346	0	0	0	0	1
Acute bronchitis	20152	0.0%	0.0191	0.137	0	0	0	0	1
Age	20152	0.0%	51.9	19.2	0	37	53	66	90
Anorexia	20152	0.0%	0.00963	0.0976	0	0	0	0	1
Asthma	20152	0.0%	0.14	0.347	0	0	0	0	1
BMI	15682	22.2%	31.3	8.36	2.08	25.6	30	35.8	93
Body temperature	3616	82.1%	37.1	0.65	29.5	36.7	37	37.3	40.2
C-reactive protein	394	98.0%	53.5	58.9	0.5	10.4	34.1	75.8	412
C-reactive protein (high sensitivity method)	155	99.2%	65.3	63.7	0.58	12.4	47.4	105	347
Cardiovascular disease	20152	0.0%	0.266	0.442	0	0	0	1	1
Chronic kidney disease	20152	0.0%	0.117	0.322	0	0	0	0	1
Chronic obstructive pulmonary disease	20152	0.0%	0.074	0.262	0	0	0	0	1
Confusion	20152	0.0%	0.0165	0.127	0	0	0	0	1
Cough	20152	0.0%	0.255	0.436	0	0	0	1	1
Diabetes	20152	0.0%	0.221	0.415	0	0	0	0	1
Diarrhea	20152	0.0%	0.0397	0.195	0	0	0	0	1
Ethnicity (Hispanic)	20152	0.0%	0.0674	0.251	0	0	0	0	1
Ethnicity (non-Hispanic)	20152	0.0%	0.39	0.488	0	0	0	1	1
Ethnicity (other)	20152	0.0%	0.416	0.493	0	0	0	1	1
Fatigue	20152	0.0%	0.0619	0.241	0	0	0	0	1
Fever	20152	0.0%	0.181	0.385	0	0	0	0	1
Gender	20152	0.0%	0.576	0.494	0	0	1	1	1
Headache	20152	0.0%	0.03	0.171	0	0	0	0	1
Hemoptysis	20152	0.0%	0.00169	0.041	0	0	0	0	1
Hospitalization (inpatient)	20152	0.0%	0.0984	0.298	0	0	0	0	1
Hypertension	20152	0.0%	0.442	0.497	0	0	0	1	1
Immunodeficiency	20152	0.0%	0.0307	0.172	0	0	0	0	1
Insurance (Medicaid)	20152	0.0%	0.00129	0.0359	0	0	0	0	1
Insurance (Medicare)	20152	0.0%	0.00372	0.0609	0	0	0	0	1
Insurance (other public)	20152	0.0%	0.0301	0.171	0	0	0	0	1
Insurance (other)	20152	0.0%	0.0274	0.163	0	0	0	0	1
Insurance (private)	20152	0.0%	0.0134	0.115	0	0	0	0	1
Insurance (selfpay)	20152	0.0%	0.00571	0.0753	0	0	0	0	1
Lactate dehydrogenase (L>P)	3	100.0%	262	69.7	186	232	277	300	323
Lactate dehydrogenase (P>L)	170	99.2%	394	322	149	256	336	425	$3.56e+03$
Lymphocytes (/100 leukocytes in blood)	1818	91.0%	23.4	11.9	0	14.7	21.9	30.4	95
Lymphocytes (/blood volume)	1808	91.0%	1.44	0.802	0.05	0.9	1.3	1.8	11
Myalgia	20152	0.0%	0.00164	0.0404	0	0	0	0	1
Neutrophils (/100 leukocytes in blood)	903	95.5%	64	16.8	0.2	56.5	66	75.5	95
Neutrophils (/blood volume)	623	96.9%	4.83	3.51	0.25	2.71	4	5.84	40.2
Nicotine dependence	20152	0.0%	0.114	0.318	0	0	0	0	1
Obesity	20152	0.0%	0.279	0.448	0	0	0	1	1
Oxygen saturation	638	96.8%	97.1	2.52	71	96	98	99	100
Paralytic syndromes	20152	0.0%	0.0175	0.131	0	0	0	0	1
Pneumonia	20152	0.0%	0.162	0.369	0	0	0	0	1
Race (African American)	20152	0.0%	0.418	0.493	0	0	0	1	1
Race (Asian)	20152	0.0%	0.0127	0.112	0	0	0	0	1
Race (Caucasian)	20152	0.0%	0.456	0.498	0	0	0	1	1
Race (multi-racial)	20152	0.0%	0.0161	0.126	0	0	0	0	1
Race (other)	20152	0.0%	0.0673	0.251	0	0	0	0	1
Rhinorrhea	20152	0.0%	0.0115	0.106	0	0	0	0	1
Shortness of breath	20152	0.0%	0.162	0.368	0	0	0	0	1
Sore throat	20152	0.0%	0.0229	0.15	0	0	0	0	1
Sputum	20152	0.0%	0.000198	0.0141	0	0	0	0	1
Vomitting	20152	0.0%	0.00695	0.0831	0	0	0	0	1
Weight	15851	21.3%	88.8	26.7	2.46	71.6	86.1	104	221

diagnosis, whether patients will enter a critical state within the next 28 days or not. In addition to demographic, clinical, and laboratory data, hospitalization and insurance types were used as predictors. Our results based on new 4031 patients unseen during training showed high predictive performance (sensitivity of 0.796 and specificity of 0.0784) and well-calibrated output probabilities. Furthermore, the interpretability analysis identified pneumonia, older age, hospitalization (inpatient), weight, shortness of breath, diabetes, race (African American), and cardiovascular disease as main predictive features risk factors and female gender and cough as risk reducing factors.

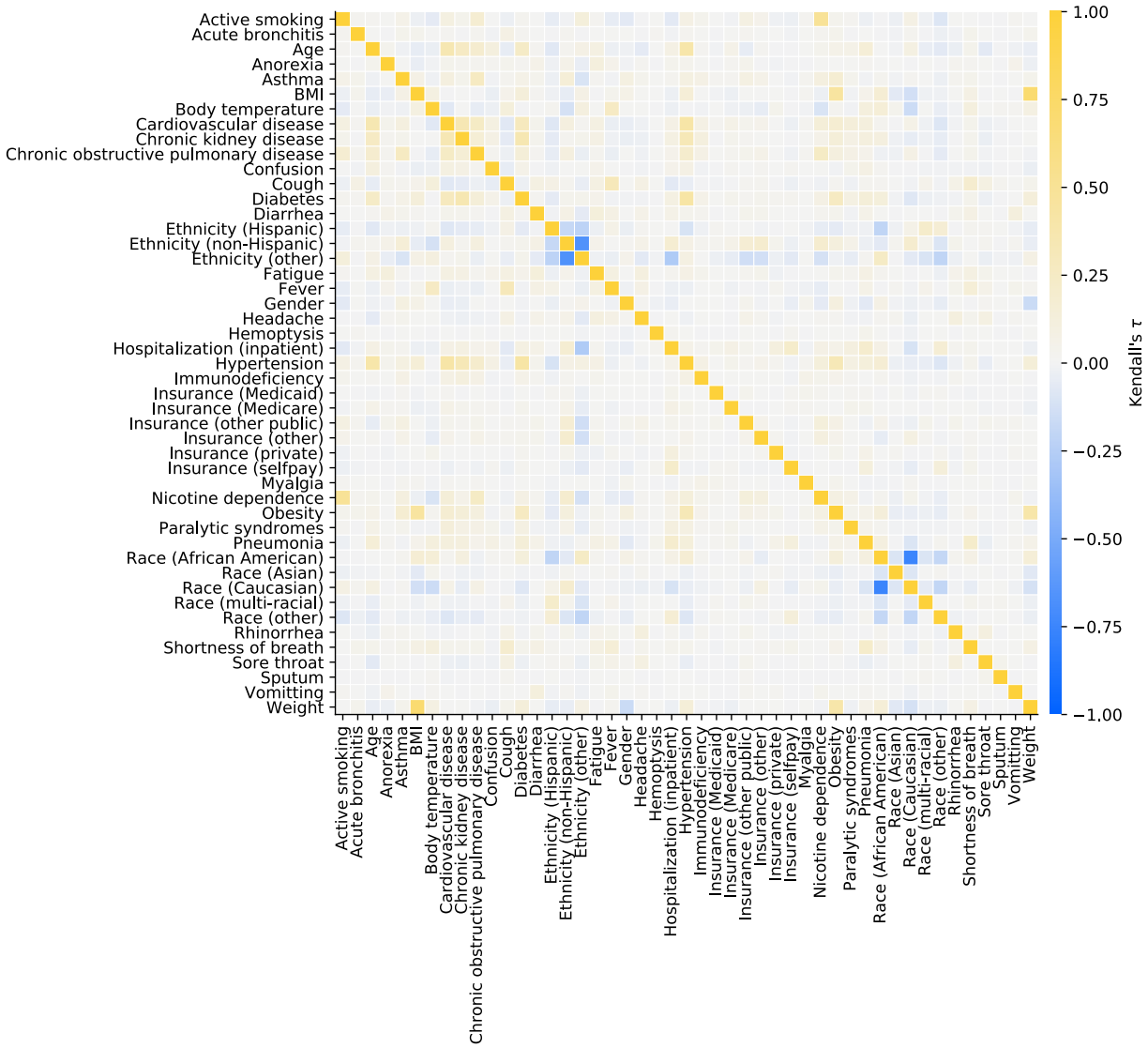


Figure 4. Feature concurvity. Kendall's τ was used to evaluate correlation between each feature combination.

4.1 Validity of the COVID-19 dataset

More than 20 000 US patients diagnosed with COVID-19 met the inclusion criteria. To the best of our knowledge, it is the largest cohort used in a retrospective analysis for predictive modeling to date based on real-world data. As highlighted in **Figure 3**, close to half of the cases were reported in Louisiana and one fourth in Ohio. This comes from the fact that the Explorys database has major contributors in the East cost of the United States. Therefore, our cohort may not be fully representative of the entire US population.

The definitions used for severe state or critical state vary across different sources (e.g., intubation prior to ICU admission, discharge to hospice, or death (Vaid et al., 2020), moderate to severe respiratory failure (Ferrari et al., 2020), oxygen requirement greater than 10 L/min or death (Haimovich et al., 2020)), or are not described in detail. Based the definition by the WHO (2020) including sepsis, septic shock, and respiratory failure (e.g., ARDS), the proportion of patients entering critical state (15.7%) in our study is within the range of prevalence (12.6% to 23.5%) reported in a review covering 21 studies (Hu et al., 2020).

Similarly, case fatality rates vary across US states and countries, as they directly depend on factors such as the number of tested people, demographics, socioeconomics, or healthcare system capacities. The death rate for the entire US is estimated to be 4.3% (Johns Hopkins University (JHU), 2020)

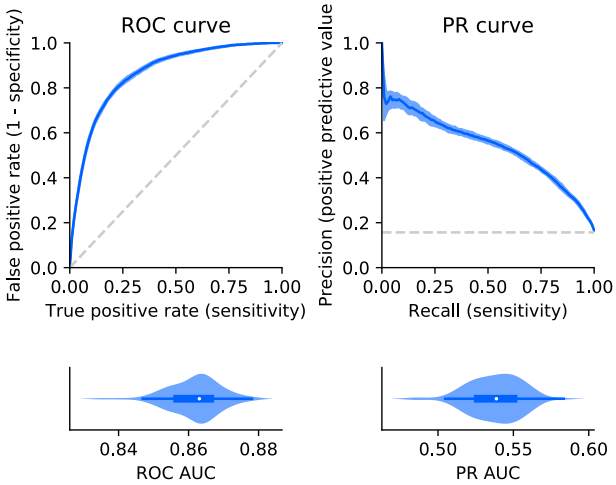


Figure 5. Model performance. Left: Receiver operating characteristic (ROC) curve and corresponding normalized violin plot of the distribution of the ROC area under the curve (AUC). Right: Precision recall (PR) curve and corresponding distribution of the PR AUC. The top plots show median and interquartile range of the performance (blue) and the chance level (no predictive value) as a reference (dashed gray line).

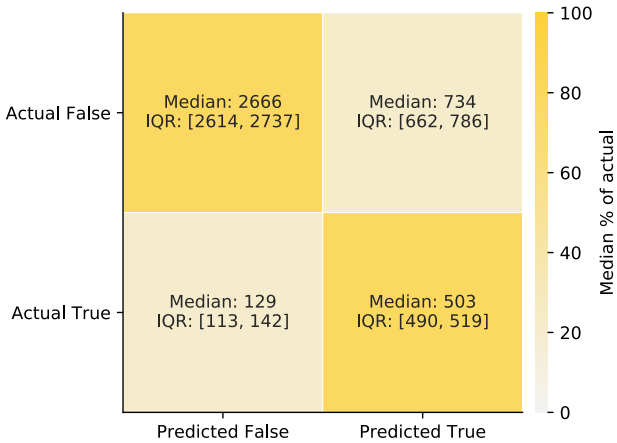


Figure 6. Confusion matrix. Confusion matrix for the predictions of the test set. True refers to entering critical state, and False refers to not entering critical state. The shades of the confusion matrix correspond to the median percentage of the actual labels (i.e., shade of the top left cell and the bottom right cell represent the median specificity and the median sensitivity, respectively).

(status July 9, 2020), while for Louisiana and in particular in New Orleans it is higher and around 4.6% (Louisiana Department of Health, 2020) and 6.5% (Johns Hopkins University (JHU), 2020), respectively. In the present work, the reported proportion of people assumed to be deceased because of COVID-19 is 5.0%. These minor difference may be

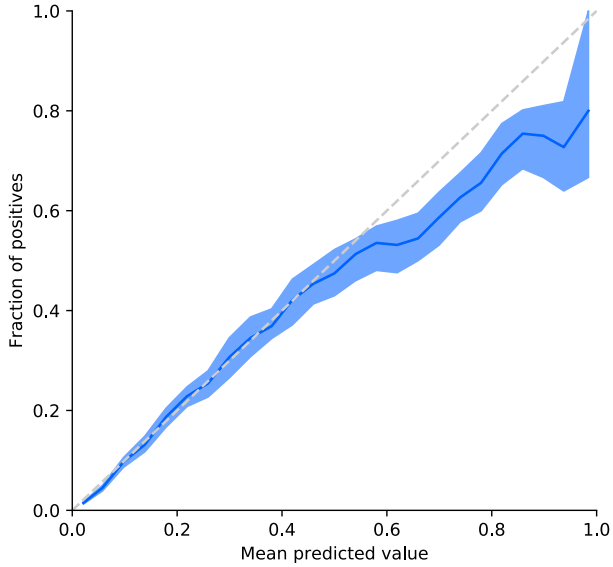


Figure 7. Model calibration. Median and interquartile range (blue) of the fraction of actual positives (labeled as critical state) for the binned mean predicted values (i.e., probabilities). The reference diagonal represents perfect calibration (dashed gray line).

justified in part by the fact that in these sources the outcome (i.e., potential death) of recently confirmed cases is yet unknown when computing the case fatality rate, hence leading to underestimation. As our analysis enforces at least 7 weeks of data after diagnosis date increasing changes of knowing the patients' outcomes, we are able to reduce this underestimation. Nevertheless, death rates based on Explorys should be cautiously interpreted, as death is not reliably reported.

Regarding demographics of our cohort, there are only minor dissimilarities to numbers reported by the Centers for Disease Control and Prevention (CDC) or US states. The interquartile of the age distribution of our cohort (37–66 years) matches 33–63 years for COVID-19 cases across the entire US (Stokes et al., 2020). The racial breakdown varies strongly across different US states. Given that Louisiana (and in particular New Orleans) is a main contributor in the Explorys network, this also explains the high proportion of African Americans. Given that Caucasian and African American represent together 87.4% of the dataset, there is a strong negative correlation between the two features, for which reason the majority group (race (Caucasian)) was considered as baseline and removed from the feature set. The proportion of female cases (57.6%) is more pronounced compared to the US-wide incidences of 406 (female) and 401 (male) cases per 100 000 persons also showing a marginally higher rate for females than males, respectively (Stokes et al., 2020). The higher count of COVID-19 cases among females (especially in African American) in the US state Georgia (Georgia

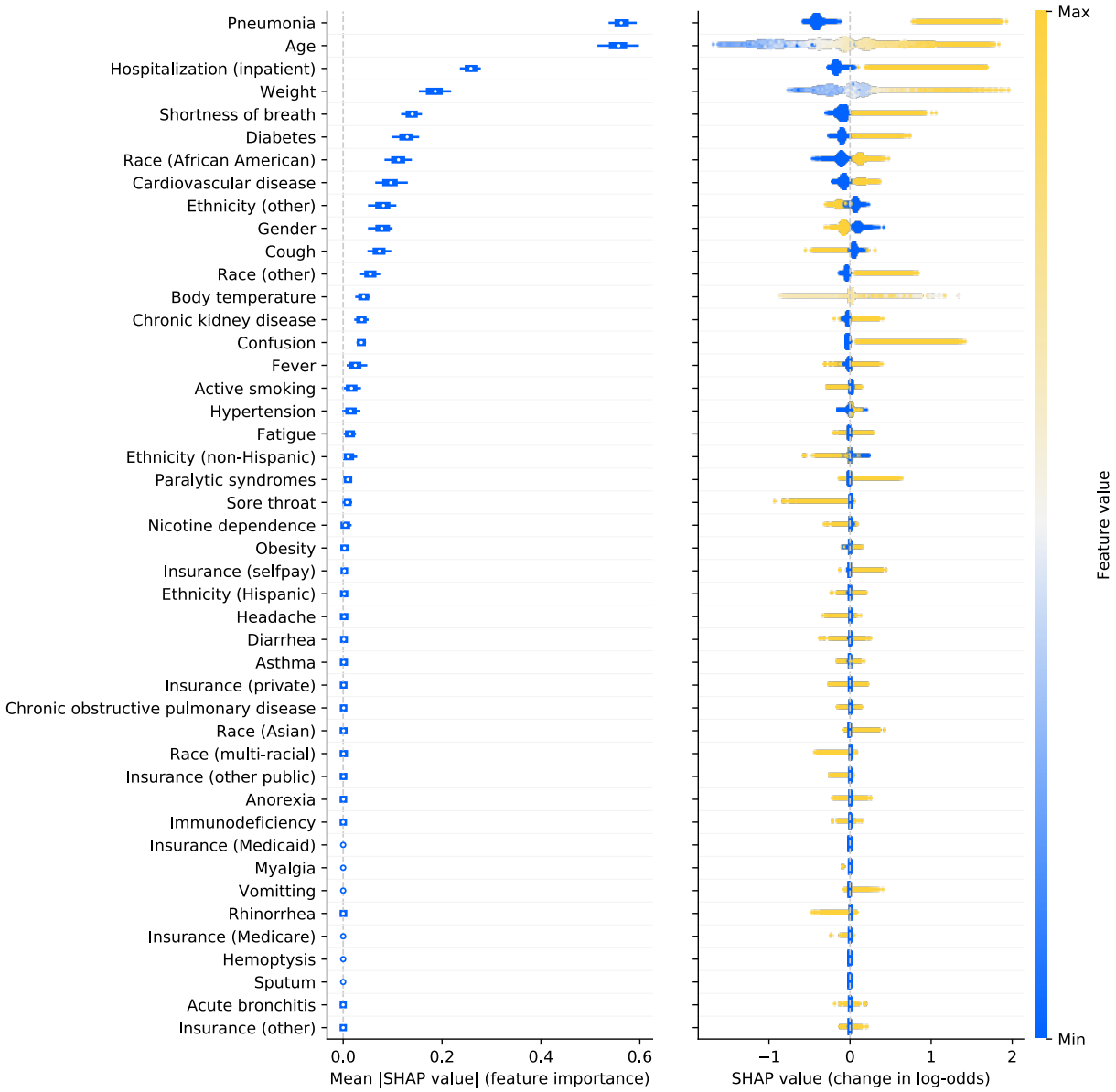


Figure 8. Model interpretability. Left: Box plots (across different seeds) of the average absolute impact of features on the model output magnitude (in log-odds) ordered by decreasing feature importance. Right: Illustration of the relation between feature values and impact (in terms of magnitude and direction) on prediction output (all seeds pooled). Each dot represents an individual patient in the test set. The color of each point corresponds to the normalized feature value (min-max normalization on test set). As an example for continuous features, older patients tend to have a higher SHAP value). For binary features, the maximum feature value 1 corresponds to presence of the feature, and 0 to absence of the feature. For gender, 1 corresponds to female.

Department of Public Health, 2020) having a similar racial distribution to Louisiana might also support our larger proportion of female cases, as African Americans cover almost half of our dataset. Since the medical system captured by Explorys is separate to the billing system, it can be expected that information on insurance types is not widely available. As a matter of fact, less than 10% of patients have a reported insurance type. Nevertheless, it has been shown to be of value (in particular knowing if the insurance is self-pay or not) for predicting critical state.

The most common underlying comorbidities identified through ICD codes in our cohort are hypertension, obesity, cardiovascular disease, diabetes, and chronic lung disease (includes asthma and chronic obstructive pulmonary disease).

As this is in line with statistics from the CDC (Stokes et al., 2020; Garg et al., 2020) as well as other studies conducted in China (e.g., Zhou et al. (2020)) and the prevalence of such features is not affected by any time window restrictions (i.e., the entire patient history was considered), it confirms the validity of the Explorys data. In contrast, all quantitative measurements identified through LOINC codes (e.g., body temperature or C-reactive protein) are missing in more than 75% of the COVID-19 cases (except for BMI and weight), and for most features it is even more than 95%. This observation could partially be explained by the fact that the acute time window only captures entries before the COVID-19 diagnosis, that lab test results are often known only after the diagnosis date, and that tests are mainly performed on hospitalized patients or patients in a severe state. Moreover, some tests might not be commonly performed in the hospitals within the Explorys network or they might just not be reported. Due to the high number of missing values, the majority of the quantitative measurements were excluded for modeling. Even though the lack of these features may compromise the performance of the model, it simplifies the model and increases its practical usability to do early predictions before extensively performing lab tests.

Since the aim of the present work is to develop a model for predictions at the time point of COVID-19 diagnosis, symptoms identified through ICD codes (e.g., fever or cough) are only extracted from the 14 days previous to the COVID-19 diagnosis. As the COVID-19 diagnosis may be early or late in the disease progression, there is the possibility to capture either early or late symptoms depending on each case. However, due to the time window restriction, the prevalence of reported symptoms tends to be lower compared to statistics including reported symptoms during the entire course of the disease (Stokes et al., 2020). Despite these lower numbers, the most common symptoms in our cohort, namely cough, fever, and shortness of breath, are confirmed by other reports and studies (Chen et al., 2020; Stokes et al., 2020; Yang et al., 2020).

Overall, the size and quality of the EHR dataset based on the Explorys database demonstrates high value with regards to demographics, chronic features, and acute symptoms in most cases, but is less suitable for laboratory test results extracted by LOINC codes. To avoid the unreliable use of quantitative features with a significant proportion of missing values, such features were removed for the prognostic modeling.

4.2 Performance

Although our dataset is based on fragmented real-world data with a high proportion of missing data, our prognostic model shows an excellent model performance in terms of ROC AUC (0.863 [0.857, 0.866]) (Mandrekar, 2010) and a substantial improvement of the PR AUC (0.539 [0.526, 0.550]) compared to chance level (0.157). Optimizing the decision threshold by maximizing the Youden's J statistic lead to a sensitivity of 0.796 [0.775, 0.821] and a specificity of 0.784 [0.769, 0.805]. Depending on the medical requirements for the prognostic model in terms of sensitivity and specificity, the threshold could easily be adjusted for a real application. As different types of datasets, inclusion/exclusion criteria, features, and prediction target definitions were used in other papers presenting the development of models predicting COVID-19 critical state, (e.g., Haimovich et al. (2020); Vaid et al. (2020), or see review by Wynants et al. (2020)), it renders it difficult to do a direct performance comparison (reported metrics were in the following ranges: ROC AUC 0.81–0.99, PR AUC 0.56–0.71, sensitivity 0.70–0.94, specificity 0.75–0.85). Furthermore, some publications do not mention metrics (e.g., PR AUC, or sensitivity and specificity) required to properly evaluate performance on an imbalanced dataset, which is the case for this type of COVID-19 prognosis. Unlike other papers (Ferrari et al., 2020; Haimovich et al., 2020; Vaid et al., 2020) usually performing a cross-validation or using a limited number of independent sets for the testing, the present approach used random, stratified train-test splits repeated 100 times to obtain a distribution of performance. Such an approach has the advantage of providing a better understanding of the generalizability model and the robustness of the performance estimate, as it is likely that a single test set might underestimate or overestimate the real performance for small testing sets. Even though our model was trained on data coming from many hospitals compared to other work being only based on a single or limited number of contributors, an external validation should be performed to better assess its generalizability.

Most publications on prognosis prediction models do not report model calibration (Wynants et al., 2020), with the exception of a few (Haimovich et al., 2020; Xie et al., 2020). The present model based on the Explorys dataset is well-calibrated, showing only minor tendency to over-forecast probabilities above 0.6. We hypothesize that this over-forecast comes from the fact that treatment features were not included in the model. Assuming that treatments reduce the probability of entering critical state, taking a treatment will lead to an overestimated probability by the model, as this information is not available to the model. In any case, over-forecast accentuating cases with relatively high probability is preferrable to under-forcast, where patients with high probability of critical case may not be identified.

Overall, our prognostic model shows excellent performance and has the advantage to provide a calibrated risk score instead of a binary classification that could potentially help healthcare professionals take better decisions to improve patients' outcome when diagnosed to COVID-19.

4.3 Model interpretability

It is not surprising to see pneumonia among the top features, as pneumonia is a diagnosis defining moderate and severe cases (WHO, 2020), which are precursor stages for critical state due to COVID-19 disease. The results from a study with 1099 patients showed that patients with severe disease had a higher incidence of physician-diagnosed pneumonia than those with nonsevere disease (Guan et al., 2020).

Increased age (e.g., above 65 years) has been confirmed by many studies to be an important risk factor for progress to grade IV and V on the pneumonia severity index and mortality of COVID-19 patients (Du et al., 2020; Liu et al., 2020b; Mehra et al., 2020). The developed model was also able to endorse existing results showing that men are, despite similar prevalence to women, more at risk for worse disease severity, independent of age (Jin et al., 2020). Similarly, obesity has been identified as a factor increasing probability of higher disease severity and lethality (Petrakis et al., 2020; Lighter et al., 2020; Guan et al., 2020). While according to our interpretability analysis the feature obesity shows marginal importance in the output of the model, the feature weight is among the top features leading to high risk (in case of high weight). It can be assumed that the feature obesity with a prevalence of 29% in our dataset compared to age-adjusted prevalence of obesity in the US is around 35% (Flegal et al., 2012) is underreported in the EHR data of our cohort. At an average US male height of 175 cm (Fryar et al., 2018), the median weight and median BMI in our dataset are very close to the threshold from overweight to obesity (BMI of $> 30\,\mathrm{kg/m^2}$). Hence it can be concluded that approximatively 50% of our patients are obese. In addition, the weight feature is a continuous variable with only 21.3% missing entries, having thus more information content and, as a result, shows higher predictive importance. A more related feature to obesity would be BMI. However, BMI was removed due to high correlation with weight and a larger proportion of missing values.

In line with the literature, the following comorbidities were also shown to drive high probabilities for critical state: diabetes (Guo et al., 2020b; Wang et al., 2020a; Yan et al., 2020b), chronic kidney disease (Cheng et al., 2020; Emami et al., 2020; Henry and Lippi, 2020), and cardiovasular diseases (Bansal, 2020; Guo et al., 2020a; Mehra et al., 2020). As a matter of fact, many elderly patients with these comorbidities use Angiotensin-converting enzyme (ACE) inhibitors and angiotensin-receptor blockers (ARBs) which upregulate the ACE-2 receptor (Zheng et al., 2020). Given that ACE-2 receptor has been proposed as a functional receptor for the cell entry mechanism of coronaviruses, it has been hypothesized that as a consequence this may lead to a higher prevalence and elevated risk for a severe disease progression after SARS-CoV-2 infection (Shahid et al., 2020).

Our model also revealed disparities in terms of probability for critical state between races: The race Caucasian showed a lower risk, while, as there is a strong negative correlation between the race feature Caucasian and African American (both together represent almost the entire cohort), the data displays that African Americans have a higher risk. This fact has been verified in several states, among others Louisiana where around 70% of deaths have occurred among African Americans, although they represent only one third of the state's population (Yancy, 2020). While a higher prevalence of comorbidities such as hypertension, diabetes, obesity, and cardiovascular disease among African Americans may be one reason for these disproportion, also late lockdowns in southern states or social determinants (e.g., living in poor areas with high housing density, high crime rates, poor access to healthy foods) may be strong contributors (Dyer, 2020; Yancy, 2020). The importance of socioeconomic factors for severe disease progression is also underlined by examining the consequences of insurance types. The SHAP analysis clearly showed that patients with self-pay healthcare tend to have a higher probability to enter critical state, as they may be reluctant to seek early medical care.

The two primary symptoms influencing the progression of the disease based on the present analysis are shortness of breath (dyspnoea) and cough, both prevalent symptoms for COVID-19 (Yang et al., 2020). Interestingly, they have opposite effects on the prediction probability of the model, with shortness of breath increasing and cough decreasing the probability for critical state. This can be explained by the fact that cough is an early symptom during mild or moderate disease, and shortness of breath develops in the late course of illness. This concurs with statistical reports from China showing higher prevalence of shortness of breath in severe cases and a higher prevalence of cough in non-severe cases and survivors (Zhao et al., 2020; Li et al., 2020; Zhou et al., 2020). Hence, if cough is reported, this may indicate that the disease in still in early stage and there is the chance that it may not lead to a critical state, whereas if shortness of breath is reported, chances for further disease progression may be much higher. Fever may be at the same time an early appearing symptom but has also been shown to be developed later during hospitalization (Guan et al., 2020; Yang et al., 2020). In addition, as reported by Zhou et al. (2020), fever has the same prevalence in survivors and non-survivors. This may also explain why it is more difficult to use it as a predictive feature, unlike for example cough, despite being also among the most prevalent symptoms (Chen et al., 2020). Nonspecific neurological symptoms like headache and confusion are less commonly reported (Chen et al., 2020). Nevertheless, confusion showed to contribute to an increase in the model's output probability. While headaches may have many potential origins not necessarily related to COVID-19, confusion may be a clearer precursor of neuroinvasion of SARS-CoV2, which has been suggested to potentially lead to respiratory failure (Asadi-Pooya and Simani, 2020).

Finally, the feature hospitalization (inpatient) also appears among others in the top features in terms of feature importance. Less than 10% of patients were already hospitalized (inpatient) in the 14 days before or at diagnosis date. The SHAP analysis showed clearly that hospitalization before COVID-19 diagnosis predicts progression towards a severe or critical state.

Overall, the findings of this work are in line with results from the vast number of studies reported in the literature and the interpretability analysis provides evidence for the validity of the prognostic prediction modeling.

4.4 Limitations

EHRs can be a powerful datasource to create evidence based on real-world data, especially when combined with a platform facilitating the structured extraction of data. However, there are trade-offs to be made when doing analyses on EHR data in contrast to the analysis of clinical study data (Kim et al., 2018). One major limitation is that patients may get diagnoses, treatments, or laboratory measurement results outside of the hospital network covered by Explorys, resulting in incomplete patient histories with potentially high proportion of missing data. For this reason, it was for example preferred to rely on COVID-19 diagnoses based on ICD-codes, instead of relying on LOINC codes for SARS-CoV-2 tests, to increase the probability of inclusion of patients being treated within the Explorys network. This highly fragmented data also requires imputation, as there is rarely a patient with a complete data record, especially when the feature set is large. The method of imputation may also introduce additional biases which are difficult to control. Moreover, features with high proportions of missing data (in particular laboratory measurement results) were removed to reduce the bias. Wherever imputation was still necessary, it was ensured that the imputation was based purely on the train set to avoid additional information leakage. While the removal of potentially important laboratory measurement results may compromise the performance of the model, it also increases practical usability of the model, as less laboratory tests are required to create a prediction. Furthermore, to ensure data privacy and prevent re-identification, patients' age is truncated, and death dates and related diagnoses and procedures are not available in Explorys. As the latter is highly relevant for the present modeling, several assumptions had to be taken. Nevertheless, resulting death rates correspond well to official COVID-19-related death rates in the US or relevant states.

An additional limitation and potential bias is linked to the data extraction using time windows. Even though the window lengths were motivated by medical reasoning, they are subject to trade-offs which is not the case for clinical studies due to precise protocols: extending the windows to capture enough information spread over multiple visits and account for delays in EHR entries, versus remaining recent enough and related to COVID-19. Furthermore, the features used in this model do not capture the time information for the individual samples (e.g., how many days before COVID-19 diagnosis the ICD code for fever entered into the system).

The model was based on US data from hospitals of the Explory network and the cohort analysis showed that the highest data contribution came from only few states, respectively counties. This resulted for example in a higher ratio of African Americans compared to the US average, it is highly likely that there are demographic and socioeconomic biases, in addition to the fact that economically disadvantaged patients may seek medical help too late. Also in terms of testing, diagnosing, and treating, the data reflects the American healthcare system.

Despite these limitations, RWE can retrospectively generate insights on a scale which would not be feasibly with an observational clinical study. Thus, it may be a starting point for subsequent, more focused clinical studies. Furthermore, approaches based on RWE might even have higher clinical applicability due to their incorporation of statistical noise while model training (Bachtiger et al., 2020).

5 Conclusions

The results of this work demonstrate that it is possible to develop an explainable machine learning model based on patient-level EHR data to predict at the time point of COVID-19 diagnosis whether individual patients will progress into critical state in the following four weeks. Without the necessity of relying on multiple laboratory test results or imaging such as CTs, this model holds promise of clinical utility due to the simplicity of the relevant features and its adequate sensitivity and specificity. Even though this prognostic model for critical state has been trained and evaluated on the largest cohort to date with over 20000 patients, it includes only cases from certain regions within the US and may therefore be biased towards sub-populations of the US and the American healthcare system. To prove its generalizability before being considered for clinical implementation, it should be validated with other datasets. This model could be augmented with treatment features (e.g., drugs or other interventions) after diagnosis in order to predict whether the respective treatments would lead to an improvement (i.e., reduction of the probability of entering critical state). Such models will never replace clinical trials to evaluate treatment effectiveness, but will help to identify responder groups or inform the design of clinical trials to eventually reduce burden on the healthcare system and optimize personalized treatment.

Disclosure/conflict-of-interest statement

MR and YK are employees of IBM Switzerland AG.

Author contributions

MR and YK lead the development of the *RWE Insights Platform*, contributed to the conception of this work, developed the methodology, implemented the use case and the modeling approach, performed the analysis, interpreted the results, and drafted the manuscript. Both authors revised the manuscript and approved the final version.

Acknowledgments

The authors would like to thank T. Egli, O. Müller, A. Peak, and S. Schumacher for contributing to the development of the *RWE Insights Platform*, and in particular T. Egli and O. Müller for their feedback on the manuscript. Further thanks go to the IBM Watson Health® team for providing access to the Explorys dataset enabling this project, B. Kolt for support and advice related to EHR data, and B. Brady for critical review of the manuscript. The *RWE Insights Platform* project is supported and sponsored by P. Bassignana and L. Böhm (IBM Switzerland AG).

References

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet* 395, 931–934. doi:10.1016/S0140-6736(20)30567-5

Armocida, B., Formenti, B., Ussai, S., Palestra, F., and Missoni, E. (2020). The italian health system and the COVID-19 challenge. *The Lancet Public Health* 5, e253. doi:10.1016/S2468-2667(20)30074-8

Asadi-Pooya, A. A. and Simani, L. (2020). Central nervous system manifestations of COVID-19: A systematic review. *Journal of the Neurological Sciences* 413, 116832

Bachtiger, P., Peters, N. S., and Walsh, S. L. (2020). Machine learning for covid-19—asking the right questions. *The Lancet Digital Health* doi:10.1016/S2589-7500(20)30162-X

Bai, X., Fang, C., Zhou, Y., Bai, S., Liu, Z., Chen, Q., Xu, Y., Xia, T., Gong, S., Xie, X., Song, D., Du, R., Zhou, C., Chen, C., Nie, D., Tu, D., Zhang, C., Liu, X., Qin, L., and Chen, W. (2020). Predicting COVID-19 malignant progression with AI techniques. *medRxiv* doi:10.1101/2020.03.20.20037325

Bansal, M. (2020). Cardiovascular disease and covid-19. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 247–250

Bullock, J., Alexandra, Luccioni, Pham, K. H., Lam, C. S. N., and Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. *arXiv* arxiv:2003.11336

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., and Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet* 395, 507–513

Cheng, Y., Luo, R., Wang, K., Zhang, M., Wang, Z., Dong, L., Li, J., Yao, Y., Ge, S., and Xu, G. (2020). Kidney disease is associated with in-hospital death of patients with COVID-19. *Kidney International* 97, 829–838

DeCaprio, D., Gartner, J., Burgess, T., Kothari, S., Sayed, S., and McCall, C. J. (2020). Building a COVID-19 vulnerability index. *arXiv* arxiv:2003.07347

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi:10.1111/j.1600-0587.2012.07348.x

Du, R.-H., Liang, L.-R., Yang, C.-Q., Wang, W., Cao, T.-Z., Li, M., Guo, G.-Y., Du, J., Zheng, C.-L., Zhu, Q., Hu, M., Li, X.-Y., Peng, P., and Shi, H.-Z. (2020). Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *European Respiratory Journal* 55. doi:10.1183/13993003.00524-2020

Dyer, O. (2020). Covid-19: Black people and other minorities are hardest hit in US. *BMJ* 369, m1483. doi:10.1136/bmj.m1483

Emami, A., Javanmardi, F., Pirbonyeh, N., and Akbari, A. (2020). Prevalence of underlying diseases in hospitalized patients with COVID-19: a systematic review and meta-analysis. *Archives of academic emergency medicine* 8, e35–e35

Feng, Z., Yu, Q., Yao, S., Luo, L., Duan, J., Yan, Z., Yang, M., Tan, H., Ma, M., Li, T., Yi, D., Mi, Z., Zhao, H., Jiang, Y., He, Z., Li, H., Nie, W., Liu, Y., Zhao, J., Luo, M., Liu, X., Rong, P., and Wang, W. (2020). Early prediction of disease progression in 2019 novel coronavirus pneumonia patients outside wuhan with CT and clinical characteristics. *medRxiv* doi:10.1101/2020.02.19.20025296

Ferrari, D., Milic, J., Tonelli, R., Ghinelli, F., Meschiari, M., Volpi, S., Faltoni, M., Franceschi, G., Iadisernia, V.,

Yaacoub, D., Ciusa, G., Bacca, E., Rogati, C., Tutone, M., Burastero, G., Raimondi, A., Menozzi, M., Franceschini, E., Cuomo, G., Corradi, L., Orlando, G., Santoro, A., Di Gaetano, M., Puzzolante, C., Carli, F., Bedini, A., Fantini, R., Tabbì, L., Castaniere, I., Busani, S., Clini, E., Girardis, M., Sarti, M., Cossarizza, A., Mussini, C., Mandreoli, F., Missier, P., and Guaraldi, G. (2020). Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia - challenges, strengths, and opportunities in a global health emergency. *medRxiv* doi: 10.1101/2020.05.30.20107888

Flegal, K. M., Carroll, M. D., Kit, B. K., and Ogden, C. L. (2012). Prevalence of Obesity and Trends in the Distribution of Body Mass Index Among US Adults, 1999-2010. *JAMA* 307, 491–497. doi:10.1001/jama.2012.39

Fryar, C. D., Kruszan-Moran, D., Gu, Q., and Ogden, C. L. (2018). Mean body weight, weight, waist circumference, and body mass index among adults: United States, 1999–2000 through 2015–2016. *National health statistics reports*

Garg, S., Kim, L., Whitaker, M., O'Halloran, A., Cummings, C., Holstein, R., Prill, M., Chai, S. J., Kirley, P. D., Alden, N. B., Kawasaki, B., Yousey-Hindes, K., Niccolai, L., Anderson, E. J., Openo, K. P., Weigel, A., Monroe, M. L., Ryan, P., Henderson, J., Kim, S., Como-Sabetti, K., Lynfield, R., Sosin, D., Torres, S., Muse, A., Bennett, N. M., Billing, L., Sutton, M., West, N., Schaffner, W., Talbot, H. K., Aquino, C., George, A., Budd, A., Brammer, L., Langley, G., Hall, A. J., and Fry, A. (2020). Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 — COVID-NET, 14 states, march 1–30, 2020. *MMWR Morb Mortal Wkly Rep* 69, 458–464. doi:10.15585/mmwr.mm6915e3

Georgia Department of Public Health (2020). Georgia COVID-19 status report https://dph.georgia.gov/covid-19-daily-status-report. Accessed on 2020-06-24

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., Cao, J., Tan, M., Xu, W., Zheng, F., Shi, Y., and Hu, B. (2020). A tool to early predict severe 2019-novel coronavirus pneumonia (COVID-19) : A multicenter study using the risk nomogram in Wuhan and Guangdong, China. *medRxiv* doi:10.1101/2020.03.17.20037515

Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., and Ziebuhr, J. (2020). The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* 5, 536–544. doi:10.1038/s41564-020-0695-z

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., and Zhong, N.-s. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine* 382, 1708–1720. doi: 10.1056/NEJMoa2002032

Guo, T., Fan, Y., Chen, M., Wu, X., Zhang, L., He, T., Wang, H., Wan, J., Wang, X., and Lu, Z. (2020a). Cardiovascular Implications of Fatal Outcomes of Patients With Coronavirus Disease 2019 (COVID-19). *JAMA Cardiology* doi: 10.1001/jamacardio.2020.1017

Guo, W., Li, M., Dong, Y., Zhou, H., Zhang, Z., Tian, C., Qin, R., Wang, H., Shen, Y., Du, K., Zhao, L., Fan, H., Luo, S., and Hu, D. (2020b). Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes/Metabolism Research and Reviews* n/a, e3319. doi:10.1002/dmrr.3319

Haimovich, A., Ravindra, N. G., Stoytchev, S., Young, H. P., Wilson, F. P., van Dijk, D., Schulz, W. L., and Taylor, R. A. (2020). Development and validation of the COVID-19 severity index (CSI): a prognostic tool for early respiratory decompensation. *medRxiv* doi:10.1101/2020.05.07.20094573

Henry, B. M. and Lippi, G. (2020). Chronic kidney disease is associated with severe coronavirus disease 2019 (covid-19) infection. *International Urology and Nephrology* 52, 1193–1194. doi:10.1007/s11255-020-02451-9

Hu, Y., Sun, J., Dai, Z., Deng, H., Li, X., Huang, Q., Wu, Y., Sun, L., and Xu, Y. (2020). Prevalence and severity of corona virus disease 2019 (COVID-19): A systematic review and meta-analysis. *Journal of Clinical Virology* 127, 104371

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., Zhang, T., Wu, Z., He, G., and Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua* 63, 537–551. doi:10.32604/cmc.2020.010691

Jin, J.-M., Bai, P., He, W., Wu, F., Liu, X.-F., Han, D.-M., Liu, S., and Yang, J.-K. (2020). Gender differences in patients with COVID-19: Focus on severity and mortality. *Frontiers in Public Health* 8, 152. doi:10.3389/fpubh.2020.00152

Johns Hopkins University (JHU) (2020). COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

https://coronavirus.jhu.edu/map.html. Accessed on 2020-07-09

Kim, H.-S., Lee, S., and Kim, J. H. (2018). Real-world evidence versus randomized controlled trial: Clinical research based on electronic medical records. *Journal of Korean medical science* 33, e213–e213. doi:10.3346/jkms.2018.33.e213

Li, K., Wu, J., Wu, F., Guo, D., Chen, L., Fang, Z., and Li, C. (2020). The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. *Investigative radiology* 55, 327–331. doi:10.1097/RLI.0000000000000672

Lighter, J., Phillips, M., Hochman, S., Sterling, S., Johnson, D., Francois, F., and Stachel, A. (2020). Obesity in patients younger than 60 years is a risk factor for covid-19 hospital admission. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* doi:10.1093/cid/ciaa415

Liu, J., Liu, Y., Xiang, P., Pu, L., Xiong, H., Li, C., Zhang, M., Tan, J., Xu, Y., Song, R., Song, M., Wang, L., Zhang, W., Han, B., Yang, L., Wang, X., Zhou, G., Zhang, T., Li, B., Wang, Y., Chen, Z., and Wang, X. (2020a). Neutrophil-to-lymphocyte ratio predicts severe illness patients with 2019 novel coronavirus in the early stage. *medRxiv* doi:10.1101/2020.02.10.20021584

Liu, K., Chen, Y., Lin, R., and Han, K. (2020b). Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection* 80, e14–e18

LOINC (2020). SARS Coronavirus 2 – LOINC https://loinc.org/sars-coronavirus-2/. Accessed on 2020-04-20

Louisiana Department of Health (2020). Louisiana Coronavirus COVID-19 http://ldh.la.gov/Coronavirus/. Accessed on 2020-07-09

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence* 2, 56–67. doi:10.1038/s42256-019-0138-9

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 5, 1315–1316

Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D., and Patel, A. N. (2020). Cardiovascular disease, drug therapy, and mortality in covid-19. *New England Journal of Medicine* 382, e102. doi:10.1056/NEJMoa2007621

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., and Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. In *Biocomputing 2018*. 192–203. doi:10.1142/9789813235533_0018

Peeri, N. C., Shrestha, N., Rahman, M. S., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W., and Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *International Journal of Epidemiology* doi:10.1093/ije/dyaa033

Petrakis, D., Margină, D., Tsarouhas, K., Tekos, F., Stan, M., Nikitovic, D., Kouretas, D., Spandidos, D. A., and Tsatsakis, A. (2020). Obesity – a risk factor for increased COVID-19 prevalence, severity and lethality (review). *Molecular medicine reports* 22, 9–19. doi:10.3892/mmr.2020.11127

Petrilli, C. M., Jones, S. A., Yang, J., Rajagopalan, H., O'Donnell, L. F., Chernyak, Y., Tobin, K., Cerfolio, R. J., Francois, F., and Horwitz, L. I. (2020). Factors associated with hospitalization and critical illness among 4,103 patients with COVID-19 disease in New York City. *medRxiv* doi:10.1101/2020.04.08.20057794

Ranney, M. L., Griffeth, V., and Jha, A. K. (2020). Critical supply shortages — the need for ventilators and personal protective equipment during the Covid-19 pandemic. *New England Journal of Medicine* 382, e41. doi:10.1056/NEJMp2006141

Shahid, Z., Kalayanamitra, R., McClafferty, B., Kepko, D., Ramgobin, D., Patel, R., Aggarwal, C. S., Vunnam, R., Sahu, N., Bhatt, D., Jones, K., Golamari, R., and Jain, R. (2020). COVID-19 and older adults: What we know. *Journal of the American Geriatrics Society* 68, 926–929. doi:10.1111/jgs.16472

Stokes, E. K., Zambrano, L. D., Anderson, K. N., Marder, E. P., Raz, K. M., Felix, S. E. B., Tie, Y., and Fullerton, K. E. (2020). Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. *MMWR Morb Mortal Wkly Rep* 69, 759–765. doi:10.15585/mmwr.mm6924e2

Vaid, A., Somani, S., Russak, A. J., De Freitas, J. K., Chaudhry, F. F., Paranjpe, I., Johnson, K. W., Lee, S. J., Miotto, R., Zhao, S., Beckmann, N., Naik, N., Arfer, K., Kia, A., Timsina, P., Lala, A., Paranjpe, M., Glowe, P., Golden, E., Danieletto, M., Singh, M., Meyer, D., O'Reilly, P. F., Huckins, L. H., Kovatch, P., Finkelstein, J., Freeman, R. M., Argulian, E., Kasarskis, A., Percha, B., Aberg, J. A., Bagiella, E., Horowitz, C. R., Murphy, B., Nestler, E. J., Schadt, E. E., Cho, J. H., Cordon-Cardo, C., Fuster, V., Charney, D. S., Reich, D. L., Bottinger, E. P., Levin,

M. A., Narula, J., Fayad, Z. A., Just, A., Charney, A. W., Nadkarni, G. N., and Glicksberg, B. S. (2020). Machine learning to predict mortality and critical events in COVID-19 positive New York City patients. *medRxiv* doi: 10.1101/2020.04.26.20073411

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC Medicine* 17, 230. doi: 10.1186/s12916-019-1466-7

Wang, B., Li, R., Lu, Z., and Huang, Y. (2020a). Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging* 12, 6049–6057

Wang, W., Tang, J., and Wei, F. (2020b). Updated understanding of the outbreak of 2019 novel coronavirus (2019-ncov) in wuhan, china. *Journal of Medical Virology* 92, 441–447. doi:10.1002/jmv.25689

Watson Health, IBM Corporation (2016). *IBM Explorys Network—Unlock the power of big data beyond the walls of your organization*. Tech. rep., Route 100, Somers, NY 10589

WHO (2020). *Severe Acute Respiratory Infections Treatment Centre*. Tech. rep., Avenue Appia 20, 1202 Geneva, Switzerland

Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L. J. M., Snell, K. I. E., Steyerberg, E. W., Wallisch, C., and van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. *BMJ* 369. doi:10.1136/bmj.m1328

Xie, J., Hungerford, D., Chen, H., Abrams, S. T., Li, S., Wang, G., Wang, Y., Kang, H., Bonnett, L., Zheng, R., Li, X., Tong, Z., Du, B., Qiu, H., and Toh, C.-H. (2020). Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv* doi:10.1101/2020.03.28.20045997

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H., and Yuan, Y. (2020a). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *medRxiv* doi:10.1101/2020.02.27.20028027

Yan, Y., Yang, Y., Wang, F., Ren, H., Zhang, S., Shi, X., Yu, X., and Dong, K. (2020b). Clinical characteristics and outcomes of patients with severe covid-19 with diabetes. *BMJ Open Diabetes Research and Care* 8. doi: 10.1136/bmjdrc-2020-001343

Yancy, C. W. (2020). COVID-19 and African Americans. *JAMA* 323, 1891–1892. doi:10.1001/jama.2020.6548

Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T., Wang, Y., Pan, S., Zou, X., Yuan, S., and Shang, Y. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine* 8, 475–481

Zhao, X., Zhang, B., Li, P., Ma, C., Gu, J., Hou, P., Guo, Z., Wu, H., and Bai, Y. (2020). Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. *medRxiv* doi: 10.1101/2020.03.17.20037572

Zheng, Y.-Y., Ma, Y.-T., Zhang, J.-Y., and Xie, X. (2020). COVID-19 and the cardiovascular system. *Nature Reviews Cardiology* 17, 259–260. doi:10.1038/s41569-020-0360-5

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* 395, 1054 – 1062. doi:https://doi.org/10.1016/S0140-6736(20)30566-3